Beta Distribution

Calvin UNIVERSITY 1876

Beta

Type

Continuous

Support

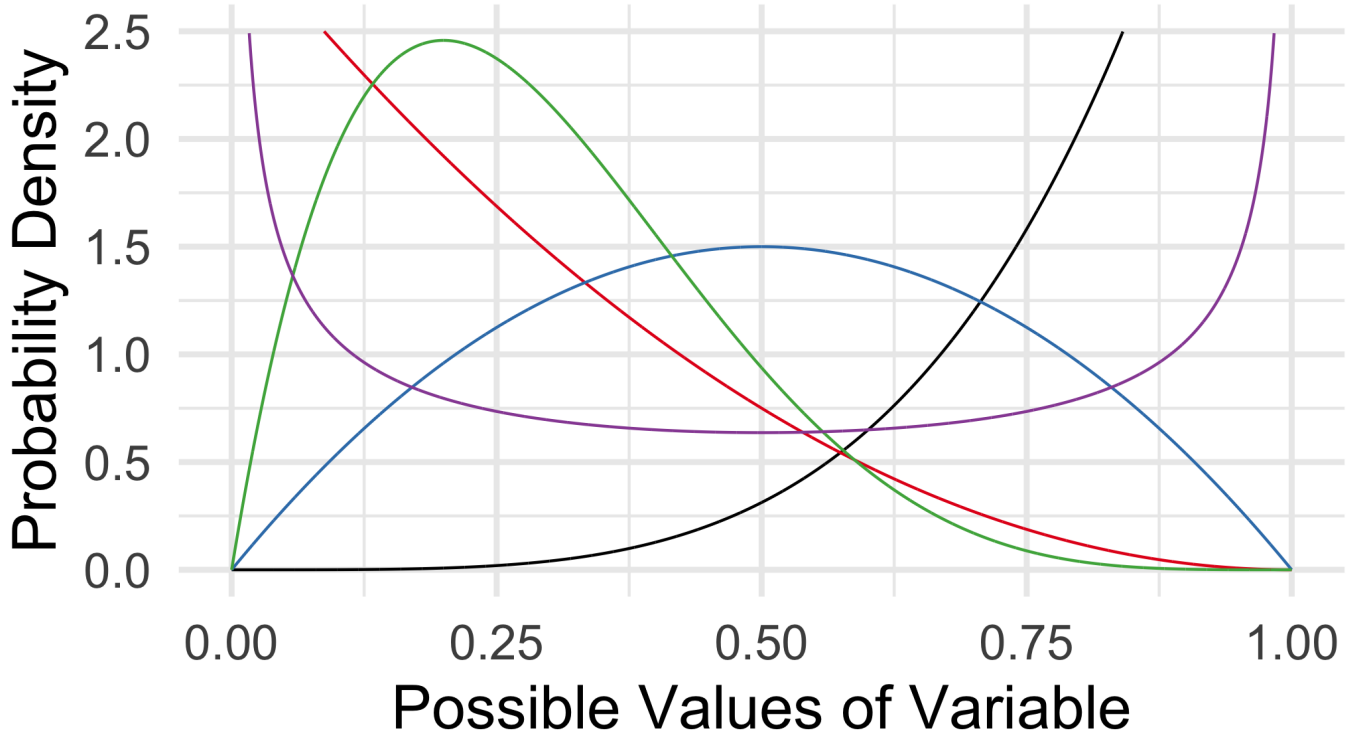
Real numbers between 0 and 1 ([0,1])

Parameters

α (shape 1) and β (shape 2), both of which must be > 1.

Shapes

This distribution can take on almost any shape, for example:



PDF or PMF

The PDF is:

$$f(x) = \frac{x^{(\alpha-1)}(1-x)^{(\beta-1)}}{B(\alpha,\beta)}$$

Where B is the Beta function (again, feel free to look up the definition if you are interested).

Examples

Beta distributions could be used for any variable that takes on values between 0-1, for example, baseball players' batting averages, or test scores (as proportions).

Binomial Distribution

Calvin UNIVERSITY 1876

Binomial

Type

Discrete

Support

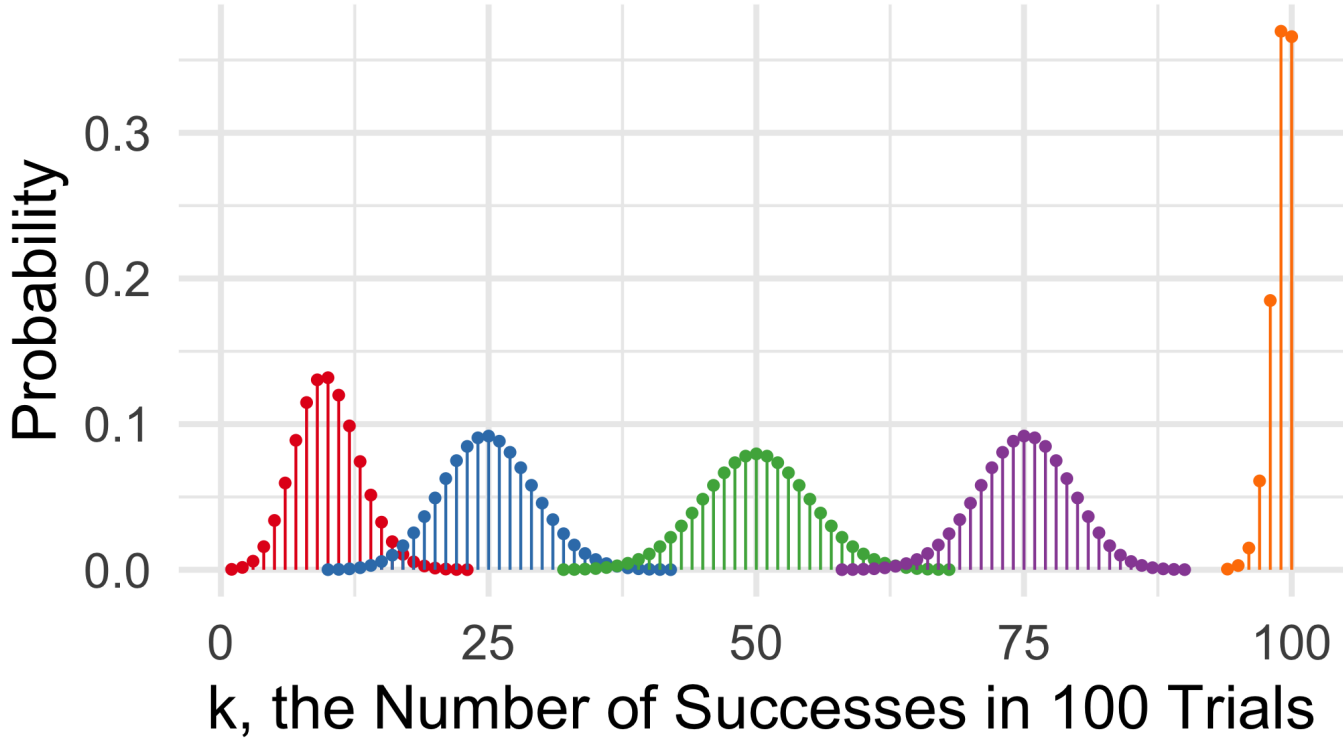
You can think about the support of this distribution two ways. Technically, the support is $k = 0, 1, 2, 3, \ldots$: 0 and positive integers, interpreted as the number of "successes" in n binomial trials. Binomial trials are independent observations of a process that has two possible outcomes, "success" or "failure", with set probabilities of each occurring. (And probabilities of success and failure must sum to 1.)

So, for each individual binomial trial, the possible outcomes are 0 and 1, often interpreted as TRUE and FALSE or "success" and "failure". For our purposes, this distribution will be useful for modelling response variables that are categorical with two possible values (and the n trials will be the n rows in our dataset).

Parameters

Parameters are n, the number of independent trials, and p, the probability of success in each trial.

The figure below shows the shape of binomial distributions with fixed $n = 100$ and varying p:



The p used were 0.1, 0.25, 0.5, 0.75, and 0.99. Can you tell which is which?

PDF or PMF

The PMF for the binomial distribution is:

$$P(X = k|n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$

Where k is the number of successes observed in n trials (you can think of k as our "x-axis variable" for this PMF).

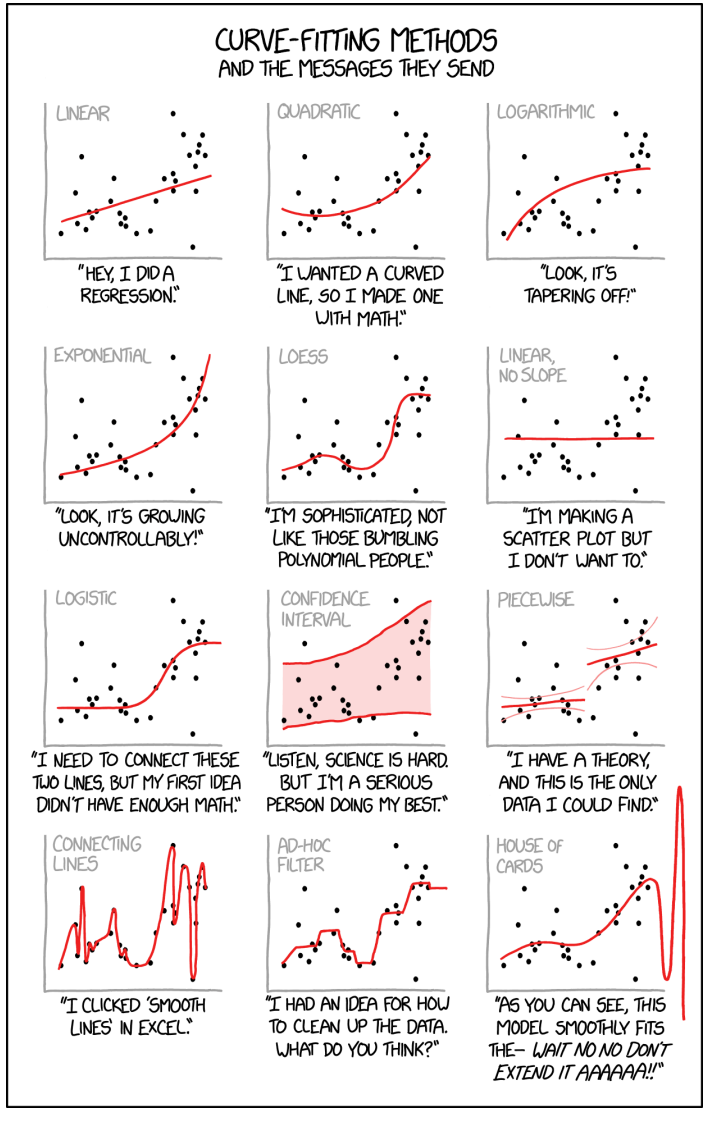
Examples

We might use this distribution to model any categorical variable with two possible values, like Age (if possible values are "adult" and "child") or health status ("has disease" or "does not have disease"). We'll think of each observation in the dataset as one of the n indpendent trials, with one of two possible outcomes for each trial.

Exponential Distribution

Calvin UNIVERSITY 1876

We won't be directly using this distribution. So just for fun...



Gamma Distribution

Calvin UNIVERSITY 1876

Gamma

Type

Continuous

Support

positive real numbers

Parameters

There are two alternate but equivalent parameterizations for the gamma distribution.

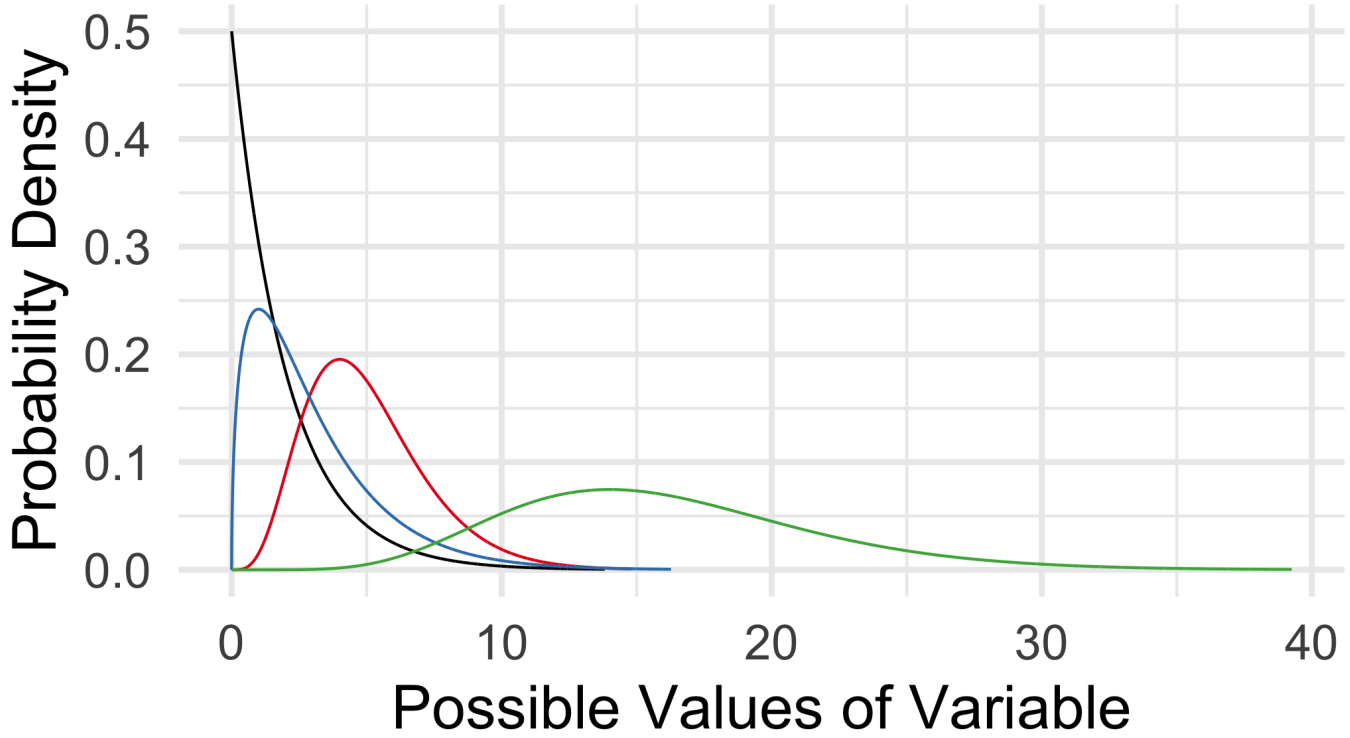
- One: α (shape) and β (rate), where $\alpha > 0$ and $\beta > 0$
- The other: k (shape) and θ (scale), where $k > 0$ and $\theta > 0$.
- Converting: $\alpha = k$ and $\beta = \frac{1}{\theta}$.

Shapes

The gamma distribution can take on a unimodal, symmetric shape or a unimodal shape with any amount of right skew (up to an exponential shape).

Note: you don't need to be familiar with exactly how the different values of parameters influence the shape of the gamma distribution PDF, so the curves here are not labelled with parameter values.

Shapes


Probability Density
0.5
0.4
0.3
0.2
0.1
0.0
0
10
20
30
40
Possible Values of Variable

PDF or PMF

The gamma distribution has PDF:

$$f(x) = \frac{1}{\Gamma(k)\theta^k} x^{(k-1)} e^{\frac{-x}{\theta}}$$

Where Γ is the Gamma function (look up the definition if you choose)

PDF or PMF

(other formulation)

$$f(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{(\alpha-1)} e^{-\beta x}$$

Examples

Gamma distributions are often used to model things like wind speed or duration of an event (any quantity that might have right skew and is never negative).

Negative Binomial Distribution

Calvin UNIVERSITY 1876

Negative Binomial

Learn about this distribution AFTER you check out the Poisson

There are two versions or "types" of this distribution, cleverly known as NB1 (type 1) and NB2 (type 2). NB1 has "constant overdispersion" -- the variance of the distribution is greater than the mean in a constant ratio. NB2 has "variable overdispersion" -- the variance is a quadratic function of the mean. The NB2 corresponds directly to conceptualization in terms of binomial trials (with the PMF giving the probability of observing y failures before the rth success). Hardin and Hilbe 2007 describe the negative binomial this way: "Instead of counts entering uniformly, we see counts entering with a specific gamma-distributed shape."

Type

Discrete

Support

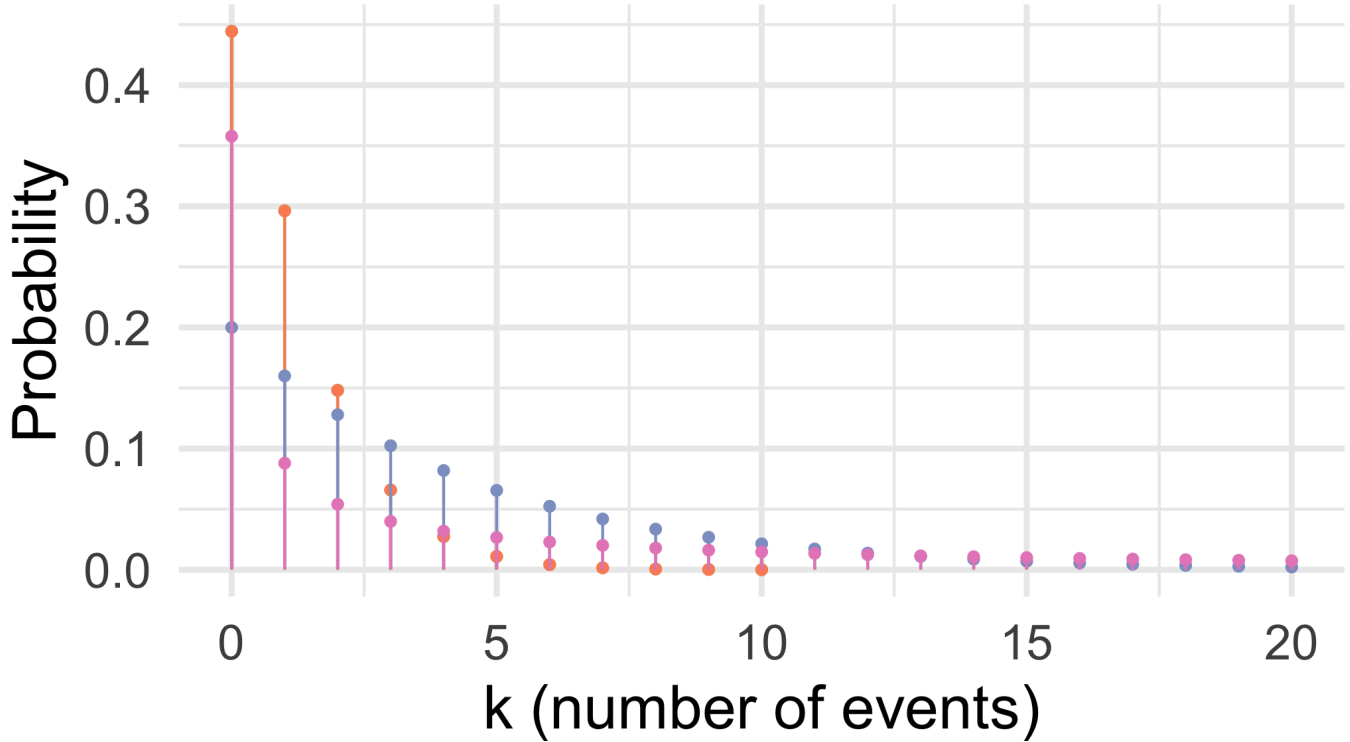
The support is 0 and positive integers (i.e., this distribution works well for count data). It also has a derivation in terms of binomial trials, but in our regression models, we will only use it with count data.

Parameters

A common parameterization of the negative binomial (online and in actuarial science) has parameters p, the probability of success on each binomial trial, and r, the number of failures observed. The PMF then gives the probability of observing k failures before the rth success in a series of Bernoulli trials.

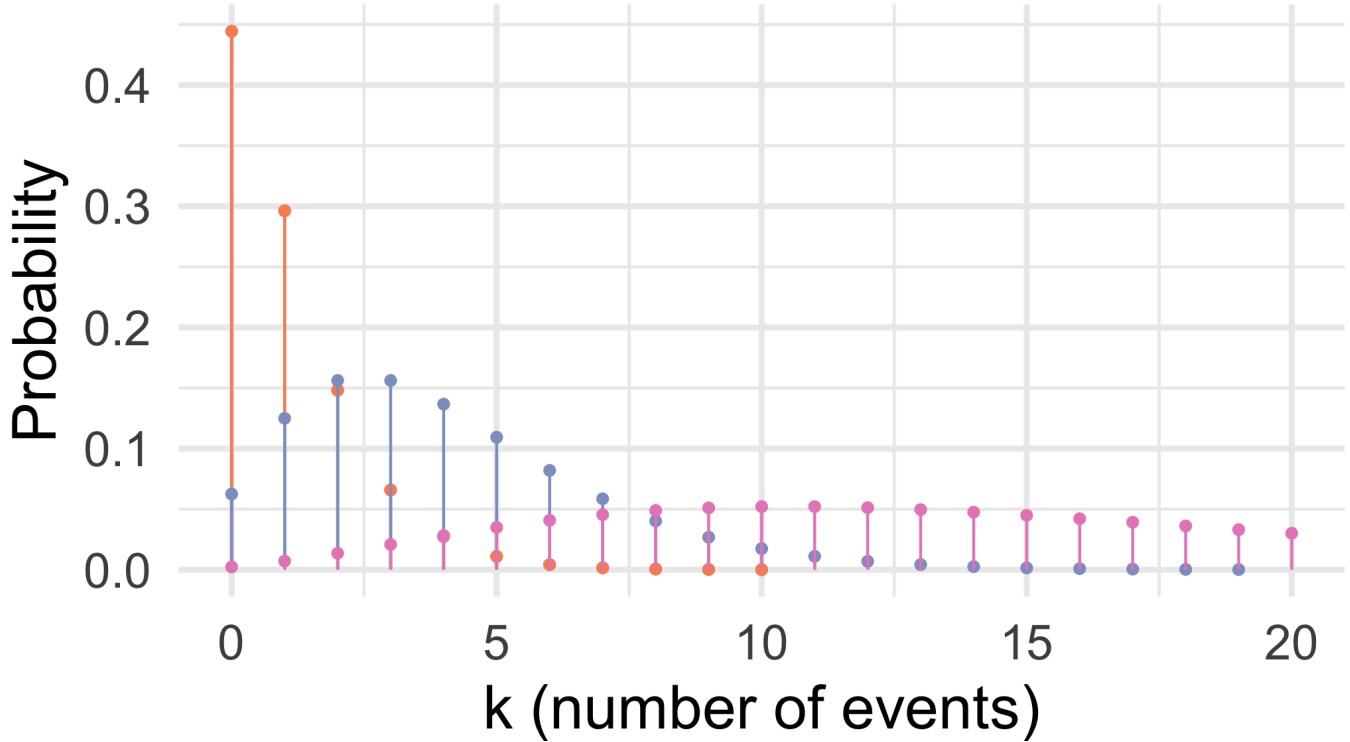
Shape, NB1

These distributions can take on unimodal shapes with varying amounts of right skew.



Shape, NB2

In NB2 (type 2) distributions the variance (spread) is larger relative to the mean.



PDF or PMF - NB1

Details of the parameterizations and likelihood and fitting of NB1 and NB2 distributions can be found in Hardin and Hilbe 2007, if you are interested. The PMF for the NB1, where the variance is a constant multiple of the mean, is:

$$f(x|\mu,\alpha) = \frac{\Gamma(x+\mu)}{\Gamma(\mu)\Gamma(x+1)}(\frac{1}{1+\alpha})^{\mu}(\frac{\alpha}{1+\alpha})^{x}$$

Where Γ is a Gamma function. Note that if α = 0 this becomes a Poisson distribution, so the Poisson is a special case of the NB1.

PDF or PMF - NB2

The PMF for the NB2, where the variance is a quadratic function of the mean, is:

$$f(x|\mu,\alpha) = \frac{\Gamma(x+\frac{1}{\alpha})}{\Gamma(\frac{1}{\alpha})\Gamma(x+1)}(\frac{1}{1+\alpha\mu})^{\frac{1}{\alpha}}(1-\frac{1}{1+\alpha\mu})^{x}$$

Examples

NB distributions are good models for overdispersed count data, where (in the regression context) the residual variance is not equal to the expected (predicted) value. (Note that if you are reading this before learning about regression models for count data, you may not understand this sentence yet...don't worry, it will make sense when you return later!) Some examples might include sightings data on numbers of animals seen on wildlife surveys, or the number of items bought per order at an online retailer.

Normal Distribution

Calvin UNIVERSITY 1876

Normal

Type

Continuous

Support

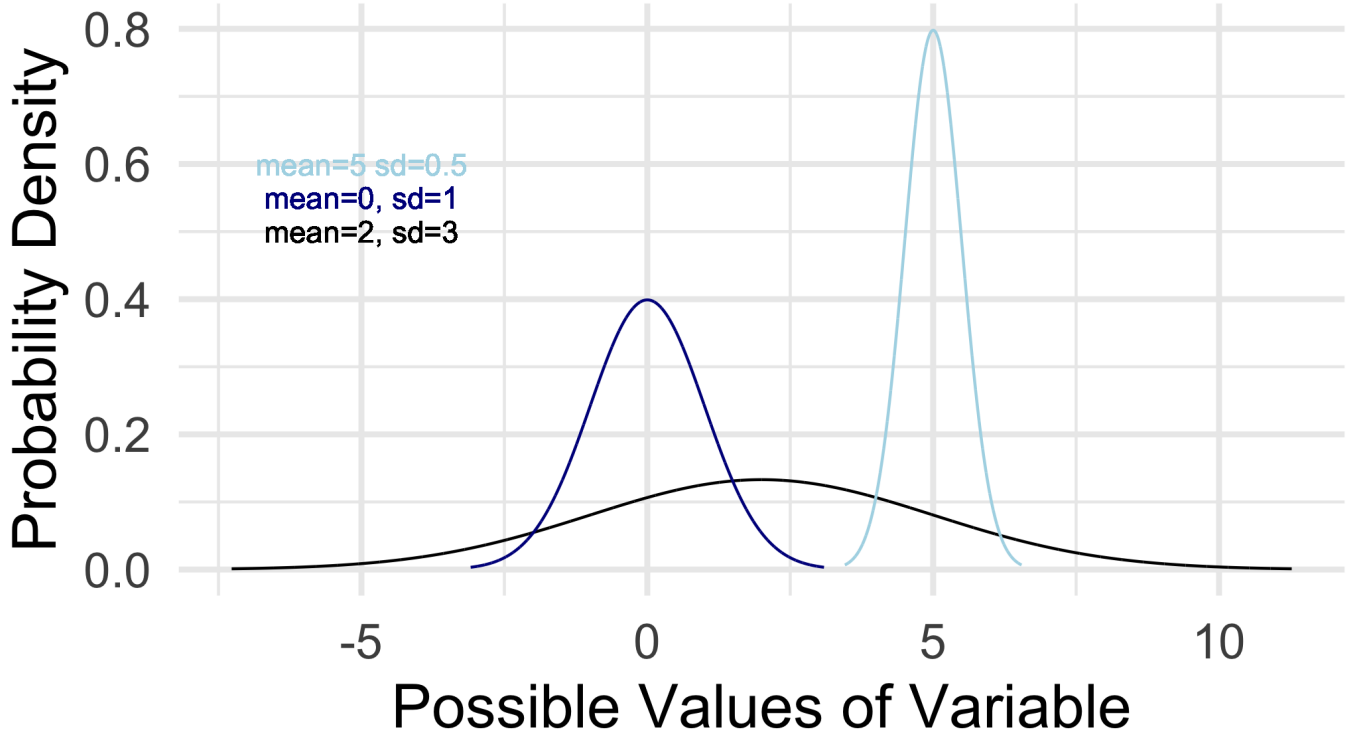
All real numbers

Parameters

- μ, the mean, which can take on any real value
- σ, the standard deviation, which can take on any positive real value

Shapes

The shape is always unimodal and symmetric.



PDF or PMF

The normal distribution has PDF:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

Examples

A normal distribution might be a good fit for data on childrens' weights in kg, or for the duration of visits at a zoo, or...

Poisson Distribution

Calvin UNIVERSITY 1876

Poisson

Type

Discrete

Support

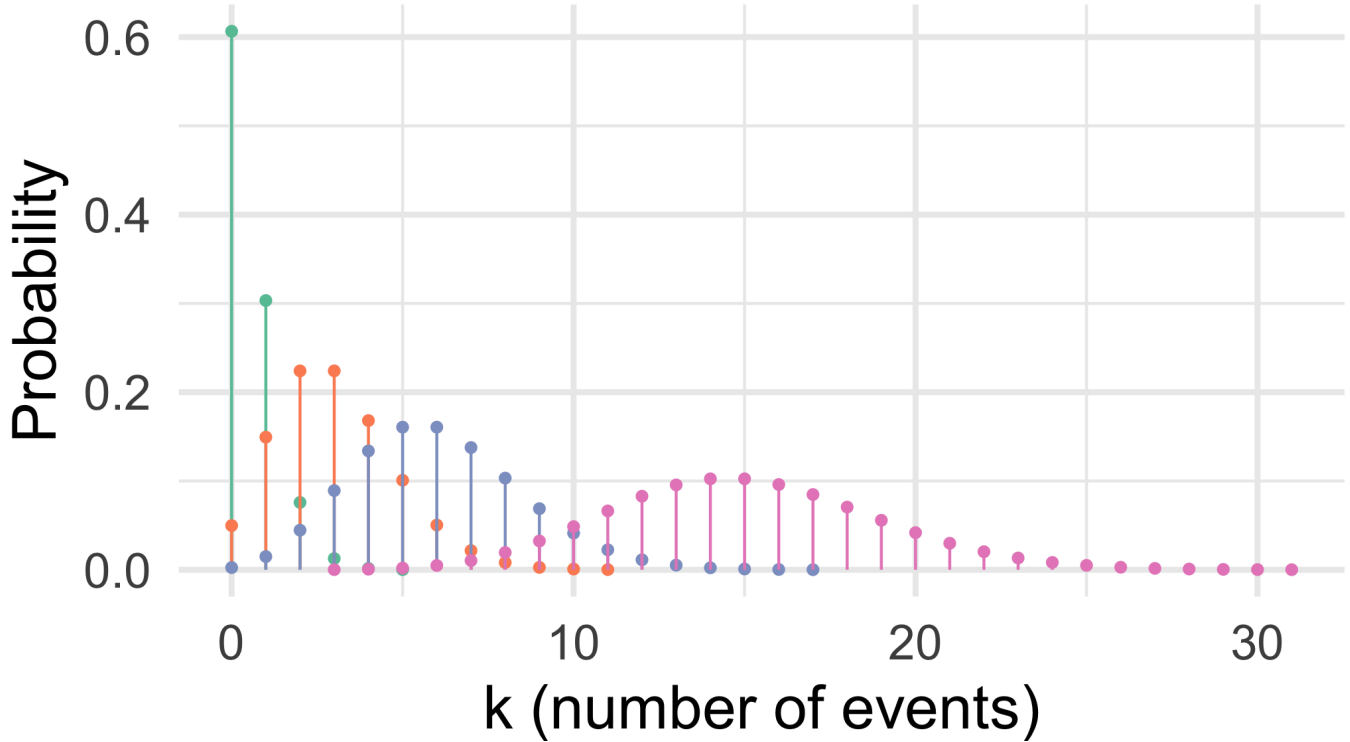
The support is 0 and positive integers (i.e., this distribution works well for count data).

Parameters

The Poisson distribution has one parameter, λ (the event rate per unit time) which must be greater than 0.

Shapes

The distribution can take on unimodal shapes with varying amounts of right skew.



PDF or PMF

The Poisson PMF is:

$$P(X = k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Examples

The Poisson distribution might be used to model any response variable that is comprised of counts, for example, the number of birds sighted in a bird survey, or the number of people admitted to an emergency room each hour. However, the mean = variance property is often a mismatch to real count data, which have variance > mean.

Tweedie Distribution

Calvin UNIVERSITY 1876

A Mixture Distribution: Tweedie

The Tweedie family of distributions is a very large one - depending on the values of the different parameters, the PMF/PDF can be written in many different ways, and it can take on many different shapes. The description below is a simplified one, geared toward the types of Tweedie distributions we are likely to try to use in regression models in this course -- mainly the "compound Poisson-gamma" type.

Some extra resources for which you will not be held responsible in this course:

- You can find an accessible description and example of this kind of distribution at: http://www.notenoughthoughts.net/posts/modeling-activity.html.
- The following site may also be useful in regard to using the Tweedie in regression models:

http://support.sas.com/documentation/cdl/en/statug/68162/HTML/default/viewer.htm#statug_genmod_details28.htm

Type

These distributions are *both* continuous and discrete - a kind of mix of a Poisson distribution and gamma distribution(s).

Support

The support is non-negative real numbers (greater than or equal to 0).

Parameters

(Note: this is one of multiple parameterizations.)

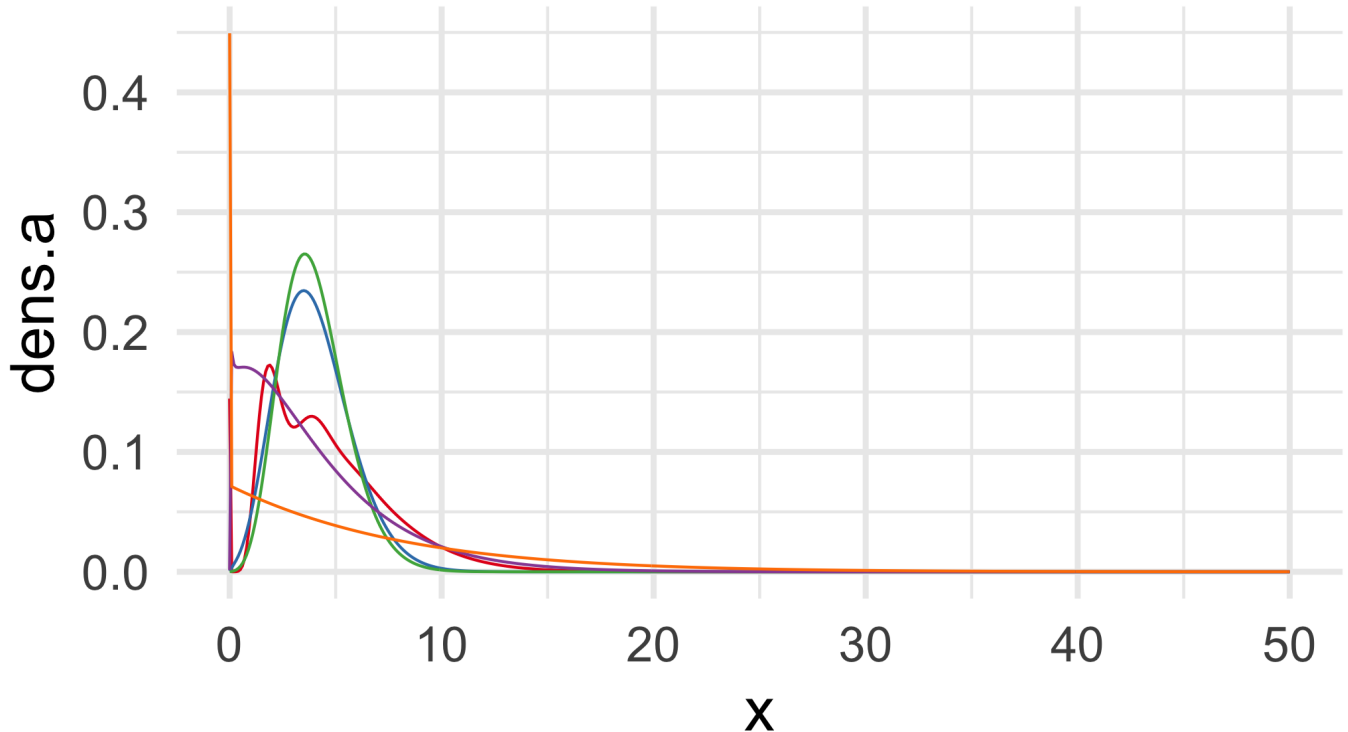
- p, the power parameter, which can be 0 (resulting in a normal distribution), 1 (Poisson distribution), $1 < p < 2$ **(a compound Poisson-gamma distribution -- what we will mainly use)**, 2 (gamma distribution), 3 (inverse Gaussian distribution), < 3 (positive stable distribution), or ∞ (extreme positive stable distribution). For compound Poisson-gamma with $1 < p < 2$, then $p = \frac{k+2}{k+1}$ where k is the parameter of the gamma distribution. When $1 < p < 2$, p closer to 1 means that the Poisson (the mass at 0) gets more "weight" in the compound distribution, and p closer to 2 means that the gamma does.

Parameters, continued

- μ. For the case where $1 < p < 2$, and the distribution is a compound of a Poisson and a gamma, then $\mu = \lambda k \theta$ where λ is the parameter of the Poisson distribution and k and θ are the parameters of the gamma.
- ϕ For the case where $1 < p < 2$, and the distribution is a compound of a Poisson and a gamma, then $\phi = \frac{\lambda^{(1-p)}(k\theta)^{(2-p)}}{2-p}$ where λ is the parameter of the Poisson distribution and k and θ are the parameters of the gamma.

Shapes

A compound Poisson-gamma Tweedie distribution can vary; a key feature is that it can have a mass at 0, then a unimodal or multimodal distribution with lots of right skew.



PDF or PMF

A Tweedie distribution with $p > 1$ has the form:

$$f(X|\mu, \phi, p) = a(x, \phi)e^{\frac{1}{\phi}(\frac{x\mu^{(1-p)}}{(1-p)} - \kappa(\mu,p))}$$

where $\kappa(\mu, p) = \frac{\mu^{(2-p)}}{(2-p)}$ if $p \neq 2$, and if $p = 2$, $\kappa(\mu, p) = log(\mu)$; but $a(x, \phi)$ is a function that does not have an analytical expression. This expression is from SAS documentation at https://support.sas.com/rnd/app/stat/examples/tweedie/tweedie.pdf.

Alternative PDF/PMF

Alternately (and more simply(?)), a Tweedie distribution with $1 < p < 2$ is a compound of a Poisson distribution with parameter λ and a gamma distribution with parameters k and θ. For example, "Suppose that airplanes arrive at an airport following a Poisson process, and the number of passengers in each airplane follows a certain [gamma] distribution. Then, the number of passengers arriving at the airport follows a compound Poisson [gamma] process $Y = \sum_{i=1}^{N} D_i$ where N is the Poisson process that the airplanes follow, and D_i is the [gamma] distribution that the passengers follow." (Thanks to D. Mao, http://math.uchicago.edu/~may/REU2013/REUPapers/Mao.pdf for this example.)

Examples

The Tweedie distributions may be useful for "zero-inflated" data, where there is a class of observations for which the observed value of the variable is always zero, and another class for which the variable takes on positive continuous values. For example, this might model the number of birds present per unit area (when the study area includes places of unsuitable habitat where none are ever found), or perhaps the quantity of alcohol consumed per week by different people (some of whom may drink varying amounts, and others of whom may never drink at all).